OPEN ACCESS

University of Dundee

University of Dundee

Dundee Discussion Papers in Economics 288

Jones, Martin

Publication date:
2015

Link to publication in Discovery Research Portal

Citation for published version (APA):
Jones, M. (2015). *Dundee Discussion Papers in Economics 288: Representations and the corruption of goods*. (Dundee Discussion Papers in Economics ; No. 288). University of Dundee.

General rights
Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Dundee Discussion Papers
in Economics

Representations and the Corruption of Goods

Martin K Jones

Department of
Economic Studies
University of Dundee
Dundee
DD1 4HN

Working Paper
No. 288
March 2015

ISSN: 1473-236X

Representations and the Corruption of Goods

Martin K. Jones[1]

Abstract: Critiques of Economics by the philosophers Michael Sandel and Ruth Grant reveal a substantive issue within economic theorising, that of attitude change, leading to the so-called "corruption of goods". It is argued that current models within economics cannot sensibly handle attitude change in the dimensions discussed by Sandel and Grant. Instead it can be modelled using *multirepresentational* games- games that allow interpretations of situations by players to change within the game. Although this may result in inconsistency within games, it is demonstrated how this can be resolved and how useful they are in analysing the corruption of goods.

1 Division of Economic Studies, School of Social Sciences, University of Dundee. DD1 4HN

Michael Sandel's book "What Money Can't Buy" (Sandel 2012) has had a considerable impact in the non-economics world as a critique of economics. It has emerged at a time when the discipline has already been under scrutiny for its failure in macroeconomics to predict the Financial Crisis of 2007. In the economics world, Sandel's impact in published articles has been limited. The only article in a major journal that confronts his arguments head-on is that of Besley (2013). Besley's article provides a good overview of Sandel's ideas and highlights recent work in economics that acts as at least a partial response to many of Sandel's criticisms. Besley admits the force of many of Sandel's arguments and recommends that economists carry out further research to tackle some of the issues that he raises. This paper aims to carry out this recommendation by analysing one of those issues.

The issue that Sandel raises, which I would like to focus upon in this paper, is that of the *corruption of goods*. This occurs when goods that have previously transacted in one institutional setting are transferred to another institutional setting. As a result of this, the norms underlying these goods are undermined leading to significant changes in behaviour and a loss in value (in the broadest sense) to the individuals involved. A similar notion is put forward by the philosopher Ruth Grant (2012), who has written a book critical of the idea of incentives as used within economics. According to Grant, the use of material incentives in a situation where incentives were not used previously can corrupt the good and lead to undesirable consequences. This is not a novel idea, as this concern has previously been brought up within economics by Hirsch (1977).

The position taken in this paper is that corruption of goods is part of a much wider unsolved problem within economics, that of changes of attitude. There have been remarkably few attempts to solve this problem, even within the narrower formulation of changes in tastes. Those attempts that have been made have generally found to be question-begging or faulty in other ways. I will argue that it is possible to model changes in attitude, and so the corruption of goods, through the use of *multirepresentational* games- games that allow for changes in players' interpretations of the world within the game.

The main argument in this paper is that current models used to explain attitude change are inadequate and that multirepresentational games provide a far better modelling technique. Multirepresentational games are presented as being similar to ordinary games except for issues relating to utilities in the games. It is assumed that, as a result of adopting different representations in a game, this may change one's assessment of an outcome's utility. The

main problem is that if one's interpretation of a situation changes partway through a game then this presents problems to a reasonable individual if they are trying to assess the overall strategy they should pursue in the game as a whole. This is because the game as whole is inconsistent when a person's utility of a given outcome changes according to their position in the game tree. This paper will look at a way in which ideas about interpretation can be introduced into games and such inconsistencies resolved without doing too much violence to the framework of conventional game theory.

Since we will be discussing attitudes and are trying to keep the widest application possible, there will be no attempt to preserve the assumption of universal self-interest within this paper. No distinction will be made between whether preferences are self-interested or social. It is assumed that agents will tend to behave in a self-interested or social way depending on the *reasons* that they have for acting in one way or another (in a manner to be explained). Given the wide range of experimental and theoretical papers that have been published asserting, and experimentally justifying, the existence of social preferences (see for example Fehr & Gachter 2000), I will not give a justification for this assumption.

Furthermore, I will assume that some of the reasons underlying choice will be based on *values*. Values will be handled informally in this paper but can be seen simply as the *importance* that an individual places on something. In the absence of universal self- interest, values become the main markers as to what an individual finds important or unimportant since one's own self-interest cannot act universally in that role. However, it will also be assumed that *sometimes* an individual's self- interest is seen as valuable and so important. In such a case one could say that the individual's self- interest is itself a value. For the purposes of this paper, it will mostly be assumed that values can be subsumed into preferences as "most preferred" items. However, these preferences can be self-interested or non-self-interested depending on the reasons[2] given.

The rejection of universal self-interest and the use of values should be central to any purported explanation of attitude change. Any purported theory that focussed on self-interested attitudes would be a highly anaemic theory that would explain very little (see Mansbridge 1998). Such a theory would fail to answer Sandel and Grant's criticisms and would be an exercise in explaining away rather than explanation. However, modern decision and game theory does not require an axiom of self-interest and neither does expected utility

[2] The role of reasons will be explained in more detail later on.

theory. All that is required is that, given the game strategies and outcomes, one's preferences are consistent.

2) Corruption of goods

To put the topic in context, we will give examples from Sandel and Grant which will provide a basis for discussion and also examples to be modelled at the end of the paper. Many of these examples are not unique to Sandel or Grant and indeed they draw heavily on ideas from economic journals, as will be indicated in the text. However, the reason for focussing on Sandel and Grant is that their notion of goods being "corrupted" is a useful generalisation that will inform the paper's focus on a new model in a way that current economic papers do not do.

Sandel's first example is the notion of a gift- why are gifts in Western countries usually given in kind rather than in monetary form? As Sandel points out, money is a better gift because it is fungible- it can be used by the recipient to buy whatever they want. However, a money gift is usually disliked in Western culture because it is seen as demonstrating a lack of care and attention by the giver. By converting gift-giving into a simple monetary transaction, one is undermining the whole value of the gift. A more extreme example is that of friendship- one cannot buy friendship because a bought friend is simply not a friend at all.

Another example is that of paying students to learn by giving them a monetary incentive to achieve high grades. The main argument in favour of this is, of course, to improve educational outcomes. However, there are a variety of arguments against this policy. One of the main arguments is the so-called crowding out argument: the monetary extrinsic motive crowds out any intrinsic motive. Essentially, education is no longer being desired for its intrinsic merits but for monetary outcomes. There is a large literature on this (See Bowles & Polania Reyes 2012 for a review) but, in essence, the outcome is that the nature of the good changes when making it into a marketable good. Instead of being something done for its own sake, education becomes a good that is traded for money.

A more extreme example is that of bribery. Bribery is the use of incentives in order to achieve an outcome favourable to the briber. An example would be to bribe a judge to let a person have a reduced sentence. As Grant points out, this actually satisfies most of the

conventional moral strictures of modern economics- bribes are non-coercive and increase welfare for both briber and recipient (assuming that they are not caught). In strict economic terms, there is an increase in welfare. The objection comes from elsewhere: bribery undermines the norm of justice that a judge is supposed to enforce. A judge therefore would be expected to reject the bribe as accepting it would be an immoral form of behaviour for him.

An example from Sandel comes from a study by Frey and Oberholzer-Gee (1996) relating to a referendum in 1993 on whether or not to have a nuclear waste repository near to the village of Wolfenschiessen in Switzerland. In a study prior to the referendum, the villagers were given a survey on whether they would vote to accept the nuclear waste repository or not. The bulk of residents indicated that they would accept it. In a subsequent survey, the villagers were asked the same question and offered an incentive of an annual monetary payment to accept the depository.

From a pure economics point of view, one would have expected the incentives to reinforce the altruistic attitude shown by the villagers but, surprisingly, the villagers, by a substantial margin, rejected the offer. In this case it is argued, civic duty is corrupted by the offer of an incentive by converting it into a market transaction. The value of accepting the good because one was fulfilling one's public duty was undermined by offering the incentive.

The final example relates to the provision of an exit option. Quite often, economists believe that if people disapprove of a particular arrangement for providing a good then they can express their disapproval by exiting from that market. As a result of this, if they stay in the market then they implicitly approve of any changes made. It follows that if people do tend to stay in the market then the good in question has not been corrupted as long as they have an exit option available and they have not used it. However, as Peter (2004) points out, this is to confuse choice with consent. One can have a low opinion of a particular way of distributing a good without wanting to exit from the distribution of the good.

An example of the above may be seen in the provision of health services in the UK's National Health Service. Over the past twenty years there has been an increased tendency for NHS services to be provided by private providers. This, it is claimed, reduces costs for the NHS and, since it is funded by tax-payers who also form the bulk of its users and the voting

population, this would seem to be beneficial[3]. However, this has been met by hostility from many quarters in spite of this. Nevertheless, most people who have the necessary money have not tried to opt out of the system and go on to another method of providing healthcare.

This paper will accept the interpretation put on these cases by Sandel and Grant as being valid. It will be granted that these are examples of the corruption of the goods or services and the norms underlying them. Rather than debating them, we will instead focus on how this corruption and the reaction to it can be modelled within economics. In other words, the aim of this paper will be to find a proper analysis of the corruption of goods as a preliminary stage towards a meaningful discussion of the idea.

3) On incentives and prosocial behaviour

Possibly the most typical attempt to explain phenomena such as these in economics is the model created by Benabou and Tirole (2006). On the face of it, this model may seem to fit the facts quite well. It explains the perverse effects of extrinsic rewards on intrinsic motivation, the effects of publicity on social behaviour, the creation of multiple norms of behaviour and the failure of social sponsors to appreciate reputational spillovers and fund social projects at the correct level. The model allows for a heterogeneous population and for multidimensional uncertainty on the part of members of that population.

Benabou and Tirole's model allows each agent to make one choice relating to their degree of involvement in a social task. The agent's choice of involvement depends on two elements: the direct benefit of involvement in the task and their reputational payoff. The former consists of the utility cost of being involved in the task C(a), the income involved, y, the level of participation $a \epsilon A$ and the preferences of the individual. These preferences are represented by two parameters v_a and v_y which represent the individual's preference for contributing to the social good and for money. It is assumed that $\mathbf{v}=(v_a,v_y)$ differs within the population according to some random distribution.

The reputational payoff depends on the posterior expectations of an agent's type **v**. This is modified for v_a by γ_a and for v_y by γ_y which represent the preferences for appearing prosocial

3 This is working on the assumption that these changes are cost reducing while maintaining the same level of service.

and disinterested respectively. These are both modified by x, the visibility or salience of one's actions. Defining $\mu_a = x\gamma_a$ and $\mu_y = x\gamma_y$ then we can summarise reputational preferences as $\boldsymbol{\mu} = (\mu_a, \mu_y)$ where $\boldsymbol{\mu}$ is also determined by a random distribution. This means that individuals in the population need to solve the following maximisation problem:

$$\max_{a \in A} \{(v_a + v_y)a - C(a) + \mu_a E(v_a|a, y) - \mu_y E(v_y|a, y)\}$$

The model can be further specified by imposing distributional assumptions onto **v** and $\boldsymbol{\mu}$, by defining C(a) more tightly and by decomposing v_a. The result is a signal extraction model where choices depend noisily on one's preferences and where different configurations give more or less convincing representations of v_a. It should be noted that the heterogeneity of agents in the model is the main source of variation and the different configurations of **v** and $\boldsymbol{\mu}$ drive most of the results.

The Benabou and Tirole model could be loosely characterised as a "behavioural" model in the sense that it uses variables that are based on behavioural, non- rational characteristics and attempts to come to conclusions that are compatible with the current state of psychological knowledge. However, in other ways, it remains fairly traditional. In particular, the main source of variation comes from the idea that agents can be split up into "types" chosen by nature (c.f. Harsanyi 1967, 1968a, 1968b). Different preferences are represented by different random selections from **v** and $\boldsymbol{\mu}$ so that each selection represents a different person. It follows that a person's preferences are "hard-wired" and that the degree of involvement in the social task depends on these fixed preferences. The various propositions in Benabou and Tirole look at the criteria at which people's behaviour changes according to their fixed preferences.

A major criticism of the Benabou and Tirole model (and related ones) is precisely that these models do not allow an individual to *change their minds*. In other words, preferences are fixed and cannot be changed. However, the situations outlined by Sandel and Grant all involve an individual swapping from one situation to another and then changing their preferences as a result of this. The Wolfenschiessen nuclear waste repository, for example, involved the same set of villagers who initially agreed to the repository but later changed their mind when an extrinsic incentive was introduced.

It may be argued that the Benabou- Tirole model could model such changes if some of the parameters were changed such as those relating to "joy of giving" u_a or "pure altruism" w_a that can be used to define v_a in public goods contexts. However, it should be noted that, while

this could indeed be done it does not provide a theory of attitude change. These parameters do not correlate to any verified and reliable psychological process, so the way in which they change when attitudes change is unknown. All that will happen is a comparative-statics change in response to an exogenous change in parameters. *Why* this happens is unknown.

It follows that the Benabou-Tirole model and others modelled in the same way cannot provide a good model for how people can change from one interpretation of a situation to another. The same follows for other models of the same provenance. Any model where the variation in the model is based on variation of types within the population is going to have difficulty explaining how the individual members of the population are going to change their minds except in an *ad hoc* fashion.

4) Other ideas on the corruption of goods

In order to deal with the points made by Sandel and Grant we will have to think more coherently about the corruption of goods as a change in attitude. It is noticeable that in the literature there are few attempts to model how attitudes change in economic contexts. There seem to be two possibilities in the literature: one put forward by Hirsch (1977) and the other put forward by Stigler and Becker (1977).

Hirsch postulated an extension of Lancaster's idea (Lancaster 1966) that goods are consumed for the sake of their characteristics rather than for being goods themselves. Under this model, there is a consumption technology that converts the consumption of goods into characteristics that the consumer wants. Utility, therefore, represents preferences over characteristics rather than goods. Hirsch postulated that this could be extended to include social norms and the environmental conditions under which they are used. In such a case, the corruption of a good would take place when these social norms were undermined or when the environment of a good is changed.

A similar, if more radical, version of this idea can be derived from the famous paper by Stigler and Becker (1977). Stigler and Becker were not concerned with the corruption of goods but rather with how to model seeming changes in tastes or preferences. Their argument was to insist that tastes should be modelled as remaining constant while all changes should be the result of changes in shadow prices. However, one can see an application of their argument to modelling corruption of goods. One way in which corruption of a good could be modelled

would be as a change in taste with respect to that good. Since this involves a change in tastes, which is unacceptable in modelling terms, it should instead be modelled as a change in a term within the utility function or, rather, a change in a term within the "production function" of each "commodity" within the utility function. A "commodity" in this case is actually a construct including all possible variables that could influence one's utility. This includes the original good, the alternatives to that good plus human capital goods related to it as well as other variables. The utility function itself does not change but variables within the "commodity" do change.

Stigler and Becker's own chosen variable for changes in the commodity production function, human capital, is not much use in this case as there is the possibility that a person may change their mind back to their original state of mind. This would imply that acquired human capital is actually lost or forgotten. However, Stigler and Becker do allow for the possibility for other inputs influencing their commodity production function so this is not fatal for the application of their theory.

Cowen (1989) has pointed out that the refusal of Stigler and Becker to allow a given utility function to change simply pushes the change back into the commodity production function or, if that is constant, into explaining exactly why one of the variables in the commodity production function varies. Stigler and Becker provide no explanation for this change. To do so would require specifying a function to explain the change in variables as well as the commodity production function. This would start an infinite regress of functions for specifying functions with no end in sight. It follows that there is no explanatory theory here; merely an endless chain of functions. Similarly Hirsch's ideas simply rely on unexplained changes in norms. It could be argued that even if this was solved it would be hopeless because attitudes are human mental states and cannot be sensibly modelled as properties of the goods themselves.

4) Attitudes, Representations and Actions

In order to establish a sound framework for examining changes of attitudes within game theory we need to understand how we are going to interpret the games. To do this I will use the ideas laid out by Rubinstein (1991) in his article on the interpretation of game theory. Rubinstein suggested that the traditional model of game theory relied on the idea of games as

a set of decisions made over physical actions defined by a set of objective rules. A "strategy" in such a context is a plan of action by a player of the game over these physical actions.

Rubinstein instead puts forward an interpretation of game theory as modelling individuals' perceptions of real life social phenomena. Games, therefore, are the game theorist's description of everything that players perceive as relevant to a specific situation. They do not necessarily conform to objective rules and, to be useful, they may not exactly reflect reality. A model should reflect the behaviour that a person would have if they were playing this particular game and the behaviour should be convergent with their behaviour in the real world situation.

A similar argument is made by Lipman (1991) in the context of trying to build a model of limited rationality. For Lipman, all games should be interpreted from the point of view of the agent so that, in the end, an agent will make his optimal choice *given* his perception of a choice situation. This means that agents can be assumed to be completely rational if one has correctly specified the choice set and preferences of the individual. However, this does not mean that the individual's perception in any way conforms to reality- perceptions could be very badly misguided and bear absolutely no correspondence to the external world at all.

One can see that there is a potentially interesting angle for explaining why a person may have differing attitudes in what physically may look like the same situation. If a person has differing perceptions of two physically similar situations then this may lead to that person having different attitudes towards the two situations since their preferences depend on those perceptions. Likewise, two people looking at the same physical situation may end up having different attitudes towards it for the same reason.

The question then becomes: where do these "perceptions" come from? Rubinstein and Lipman are more concerned with modelling aspects of game theory than going into this question in any depth. Rubinstein in his 1991 paper[4], for example, briefly speculates that the logic of perception is best examined by evolutionary biology and rejects analysis involving game theory but doesn't give any arguments to support his case. I would argue that this is too hasty. Our perceptions of the social world inevitably involve non-physical elements such as

[4] As is discussed later, Rubinstein goes into more detail in his 1998 book.

norms, conventions, ethics and institutions that can be discovered via our perception but which are not studied by evolutionary biology.

I would argue that the logical analogy of Rubinstein and Lipman's notion of perception in psychology and cognitive science is that of the *mental representation*. A mental representation is an interpretation of the world held internally by a human being. This is an analogy, used in cognitive science, to representations held by information processing devices such as computers. Within cognitive science, representations are intimately connected with actions and it will be argued that this is crucial in formulating a theory of attitude change.

It is assumed that the actions we are concerned with are intentional actions i.e. actions that are caused by the reasons for those actions. These reasons for actions have a belief component and a desire component (Davidson 1963). It will be assumed that these reasons can be converted into preferences where the belief component can be measured by subjective probabilities while the desire component can be measured by utilities. This conforms with conventional expected utility theory[5,6].

The next stage of the argument is due to Fodor (1987) who highlighted the ubiquity of so-called folk psychology in our reasoning. Folk psychology is the tendency to attribute an action by another person to mental states in that person. In other words, if a person carries out an intentional action then we tend to believe that it is because they have a reason to do it. If this is correct then we have to accept that there is a mental state behind every action which is a psychological disposition towards a specific content. So, for example, if a person playing cricket swings his bat to hit a ball then onlookers assume that this is intentional and that he is attempting to score runs. The psychological disposition here is the determination of the player to score runs, while the content is the concept of scoring runs.

Any mental state that has content is a representation and it is Fodor's main contention that any intentional causation of an action *must* be accompanied by an explicit representation in the mind. As Fodor points out, this links in very closely with the cognitive picture of the mind. The cognitive picture of the mind makes a close analogy between the human mind and a computer. In order for a computer to operate, it needs a representation of the data from the external world for purposes of storage and processing. A human mind, by this analogy,

[5] The intentional part is important here. If an action is unintentional then there is not necessarily any reason attached to it. This would be the case if someone did something accidentally.
[6] In this paper I will ignore any issues associated with measurement of preferences.

requires exactly the same thing. Fodor notes that this analogy is probably the only one available and it is the one that fits in best with the folk psychological picture.

Fodor tends to restrict the notion of representations to beliefs on the grounds that representations are about content and that content tends to be about what is the case. However Smith (1987) has pointed out that one need not restrict representations so tightly since content could be about how the world *should* be rather than just about how it is. In other words, one could be motivated to an action by one's goals, which would involve rearranging the world to fit what one wants to happen. In order to do this, one would need a representation of one's goals and desires.

This all suggests that intentional actions are caused by reasons which split up into beliefs and desires as Davidson claims and this can be modelled by expected utility theory. However, we can go further and posit that our reasons for actions are bound up in mental representations that include both belief elements that show the world as it is and desire elements that represent the world as one wants it to be. A human being can potentially entertain a wide variety of representations, including many different representations of the same situation. We will assume that, given a choice set comprising a given number of possible actions, we can amalgamate the representations of each feasible action together into one consistent representation for the choice set. From now on, when we refer to "representations" we will refer to representations over a whole choice set and if we want to refer to content associated with individual actions we will simply refer to them as "reasons".

Since we are focussing on the effects of representations on attitudes, we will concentrate on changes to the payoffs and assume that the structure of the game, apart from payoffs, remains the same as before. It follows that a representation can be implemented in a game as an allocation of expected utilities to each player's choice set. This may seem trivial since expected utilities are allocated to games all the time. However, it should be noticed that, unlike conventional decision theory, there is more than one possible representation for each choice set. It may be possible for the same choice set to have completely different allocations of utilities according to the representation selected.

For the purposes of this paper we will assume that the payoffs in a game $\mathbb{G}$ are represented by utilities attached to the outcomes s_u for all u. Representations in games are assumed to be determined by the *perceived attributes* of a situation (c.f. Keeney & Raiffa 1976). This means that, in each situation, a player is able to subjectively perceive it to have

certain qualities that can be measured by an external observer as attributes. It will be assumed that each representation will be determined by a vector of attributes which will be labelled **x** (or **y** for an alternative representation). Different representations can be formed by particular values of each attribute in the attribute vector.

Each attribute is assumed to be monotonic with the utility function that determines the preferences of that attribute. If this is not the case then it is assumed that the attribute can be reformulated so that it is indeed monotonic (See Keeney & Raiffa ch. 2). Furthermore, it is also assumed, for simplicity, that the utility functions with attributes as arguments are twice differentiable in those attributes. Given the role of interpretation in forming representations and the subjectivity of attributes, it will be assumed that attribute vectors do not overlap. Even if an objective measure forms the basis for two subjective attributes in two different attribute vectors, it will be assumed that these count as separate attributes because of the potentially different interpretations applied to them.

Attributes by themselves do not reflect the content and nature of representations in the cognitive science literature. This nature is inherently propositional (see Fodor 1987) so that different psychological attitudes are expressed as *propositional attitudes*. It would be useful to incorporate these ideas into the payoffs of games in order to explore the logical relationships between utilities and propositional attitudes. In order to do this, we will impose an extra layer of structure into utility functions in the form of *predicates*[7]. It is assumed that every value h, of attribute x_n in the attribute vector **x**, has a corresponding simple predicate $a(x_{hn})$ where x_{hn} is a value of $x_n \in$ **x**. These simple predicates simply describe the value of the attribute. From these simple predicates more complex predicates can be built up by using logical functions and relations. $H(\mathbf{x}) \in И(\mathbf{x})$ is a *representation* which is a predicate consisting of a concatenation of all the complex predicates held by a player with relation to the situation covered by the attribute vector **x**. $И(\mathbf{x})$ represents the set of possible concatenated complex predicates that can be constructed with different values of the attribute vector **x**. A similar representation could be constructed for attribute vector **y** so that the representation is $E(\mathbf{y}) \in Д(\mathbf{y})$. It is assumed that there are no complex predicates that have attributes from more than one attribute vector.

[7] A similar idea is explored in Weirich (2010) where proposition-based utilities are used to investigate framing issues.

One major assumption that will be made is that complex predicates should not be self-contradictory nor should they contradict other complex predicates within the same representation. It should also be noted that the restrictions placed on attributes implicitly place some quite tight assumptions on the complex predicates that can be formed. To take an example, a complex predicate should not end up negating the direction of an attribute so that the attribute is no longer monotonic with the utility.

Assume that each game $\mathbb{G}$ has an associated set of possible predicates denoted $\mathcal{V}$. This set represents all possible complex predicates that could be constructed from all possible attribute vectors in set **X** that could be associated with the game. It follows that all H(**x**), E(**y**) ϵ $\mathcal{V}$ and **x**,**y** ϵ**X**. Imagine that a set of actions in the game has resulted in an outcome s_k. Suppose $M_k \subset \mathcal{V}$ is the set of possible complex predicates for each s_k while $\mathbf{X_k} \subset \mathbf{X}$ is the corresponding set of attribute vectors. The complex predicate $q_k \epsilon M_k$ is a disjunction of possible predicates that potentially could hold once s_k has been selected in the game[8]. Different complex predicates q_k for each option are formed depending on the values of the attributes in $\mathbf{X_k}$. It is assumed that H(**x**) and q_k overlap but are not subsets of each other[9]. In essence, representations operate by imposing restrictions on the predicates q_k for each outcome s_k.

Finally, we assume that representations in the form of complex predicates can be incorporated into utilities and act as part of the structure of the utility function, translating the attributes into preferences. It will be assumed that individuals have preferences between predicates and that these are representable by a utility function. Each representation essentially creates its own utility function. A different representation over the same set of attributes (and the same external utility function) would result in a different set of utilities. This operationalises the idea that representations are essentially an allocation of utilities to outcomes.

In this paper we will be looking at utilities in multidimensional terms (see Weirich 2001 and Broome 1991). This essentially means that we will look at two dimensions across

[8] Note that, within q_k many of the predicates may be contradictory. However, this is not a problem as q_k is simply a partitioning of *possible* predicates that could be taken up by the player when they arrive at outcome s_i. It implies nothing about the consistency of the predicates that the players actually hold in their representations.

[9] This is because a representation may imply other options apart from q_k and q_k represents all possible predicates in an option while A(**x**) is just concerned with those possible predicates actually held by the player.

utilities- one across outcomes and one across attitudes. Since we are using attributes, we can use some of the concepts associated with multiattribute utility, such as utility independence, to govern relationships between utilities defined over different outcomes (or different interpretations of the same outcome). However, it will be assumed that the utilities themselves can be decomposed using preferential independence assumptions that hold between attitude propositions. Hence, utilities can be decomposed the same way and the parts compared with the similar parts of other decomposed utilities.

We will start with the complex predicates that form representations. The predicates $H_g(\mathbf{z}) \in \mathit{И}(\mathbf{x})$ for all g and $J(\mathbf{y}) \in Д(\mathbf{y})$ are different representations over attributes **z** and **y**.

Definition 1: Predicates $H_g(\mathbf{z})$ are preferentially independent of any predicate $J(\mathbf{y})$ if the conditional preferences between different $H_g(\mathbf{z})$ predicates given $J(\mathbf{y})$ do not depend on $J(\mathbf{y})$[10].

However there are also the complex predicates across multiple attribute vectors formed by choices between options.

Definition 2: Predicates $q_k \in M_k$ are preferentially independent of any predicate $q_g \notin M_k$ if the conditional preferences between different predicates $q_k \in M_k$ given $q_g \notin M_k$ do not depend on $q_g \notin M_k$.

Various aspects of the game and its representations can be said to be preferentially independent of each other. It will be assumed, for example, that, for all k, q_k and its complement $\neg q_k$ are both preferentially independent of each other. This fits in with the intuition that preferences for an outcome should not be influenced by aspects of the problem outside the outcome and *vice versa*. Likewise, a representation $H(\mathbf{x})$ is assumed to be preferentially independent of $\neg H(\mathbf{x})$ since the preferences between representations should not be influenced by external factors.

5) The Selection of Representations

One question that remains is how one representation is selected over another. How is it that one comes to allocate one set of utilities to outcomes in a choice set rather than

[10] One could also define a complex predicate over two attribute vectors to be preferentially independent of a third in a similar way- this will be useful later on.

another? Call the representation that is actually used to allocate the utilities in a choice set the *active representation*, labelled A(**x**) ϵ И(**x**) for an attribute vector **x**. There will be alternative representations of which we will focus on one, labelled D(**y**) ϵ Д(**y**) for attribute vector **y**. One issue is how many representations will be available for selection as the active representation. At first it might be thought that this is an impossible task to decide. There are, potentially, an infinite number of ways in which one can interpret a particular situation, even given a limited number of attributes, and one might be tempted to say that we are only constrained by our imagination.

However, there are a variety of practical constraints. Firstly, we are boundedly rational, with limited time for formulating alternative representations in our mind. Our attention time is rationed and we have limited processing time for some of the more complex ideas. Secondly, we tend to look to other people for many of our ideas and we tend to adopt representations that have been formulated elsewhere. These tend to be limited in number because they have to be easily transmitted between individuals (see Sperber 1996 for a detailed analysis of this). Thirdly, we will tend to focus on those representations that fit in with our values and so reflect our priorities. Finally, representations have to make sense in terms of one's own experience and knowledge. Given this, one would expect representations to be discrete from each other rather than a continuity. Arbitrarily changing the value of an attribute in a representation may make the whole representation incoherent either internally or with the external world. Within these constraints, it is likely that the number of representations available for choosing will be comparatively small.

In such a situation, it would be necessary to decide which should be the active representation for the individual. From the individual's point of view, this would mean that she has to select one of these representations as being the "best" in some sense and then use the representation to assign utilities to perceived moves in the game. In other words, we have to find a *selection mechanism* for representations. This search is made hard by the fact that we still know comparatively little about how the mind works. However, there are many possibilities that can be at least partially rejected on general principles.

One of these is that the selection mechanism is essentially unconscious i.e. we do not select representations through an exercise of will but by some unknown process in our mind of which we are not aware. This possibility seems very attractive because, sometimes, we do indeed seem to assume one particular interpretation of the world without actually thinking

about it. The existence of habits and inherited taboos that persist in the face of alternative ideas is sufficient evidence for this.

However, this surely cannot cover all cases. There are many times when one considers one's position consciously and tries to understand what is going on. Indeed, one could *define* "understanding" in this manner: the fitting of a representation to a particular situation. Given that trying to understand something is a conscious action, it follows that at least some of the situations where an individual selects representations must be the result of conscious processes. Furthermore, one could argue that any action that is unconscious cannot be modelled and cannot be analysed as a change in mind.

Another possibility is that the selection mechanism is in some sense automatic or unintentional. This would mean that there is a conscious process whereby the mind would select one representation over another. In a sense, this must have some truth in it. Visual perception and the understanding of visual perception by the brain are, by and large, processes that we do consciously but that we do not deliberately control. However, the results of visual or other sense perception cannot be the only content of representations. For example, in social situations there are elements that simply cannot be perceived through the senses but must be understood by use of our reasoning powers. Examples of this are social institutions, norms, conventions, taboos etc. None of these can be perceived directly but must be *deduced* from visual or other signs.

Given that at least some selection mechanisms, including most involved in social situations, must be the result of intentional reasoning, one possibility is that representations are best selected by a comparison of the beliefs in a given representation. This would mean that the representation which one judges to have the highest subjective probability would be the one that is taken on by the individual. It is certainly true that belief forms a major part in the selection of representations. One such model that has been developed is that of Rubinstein (1998) who builds a model of perceptrons based around the idea that one's beliefs are limited and that one only gains a partial picture of the external world. Rubinstein sees beliefs as being actively formed by agents as part of an optimisation process in which agents select their knowledge partition, given their own constraints on information processing. This model is quite limited, in that it focusses on rougher information partitions imposed on a continuous variable, but it shows how such models could be developed.

However, this cannot be the whole story. It ignores the role of wishful thinking and self-deception in human affairs i.e. the fact that sometimes the content of one's representation is more consistent with what one wants to be the case than what is actually the case. It also ignores situations where evidence for any representation is lacking but there is an overwhelming sense of the importance of a particular representation. Also not taken into account is the fact that people judge information according to *values* such as objectivity and accuracy (Williams 2002). In general, more accurate or objective information is judged to be more important i.e. is more highly valued than inaccurate or non-objective information. Values tend to enter into one's desires rather than beliefs so in all cases one's desires have an influential role in selecting representations.

The proposal that I wish to make in this paper is that the best way of modelling the selection mechanism is as the result of an autonomous choice made by the agent. In other words, representations are chosen in much the same way that one chooses goods. Human beings are self-aware and are able to analyse their own beliefs and desires to see whether they fit into a given situation. In the process of this analysis they need to make a judgement as to which representation best fits the situation and to *choose* between those representations that are available. While there are some restrictions on how we can model these choices, I believe that there is no intrinsic problem with this method.

I would contend that this is not an unusual way of thinking. Whenever we go into a new situation we are always looking for clues as to how we *should* behave and also what sort of situation it is. If we visit a foreign country we try to find out and understand the social structures, institutions and mores of that country so that we can accommodate them and not cause embarrassment. We then work out what our understanding is of the culture and adapt to it i.e. we choose the representation that conforms most to our beliefs and desires. Although this may seem plausible in some respects, there are some objections that may be made against these ideas. I will first of all explain the theory in more detail and then focus on some of the objections in the discussion section.

One necessary concept for such a theory is that of a *mental action* (Proust 2001, Geach 1957, O'Brien & Soteriou 2009)[11]. A mental action is an intentional action by the mind that has as its goal another mental process. Examples of this are easy to find. A person tries to remember where his keys are. A student tries to concentrate in a lecture. An

[11] I fact the notion of a mental action, as described here, was first implicitly analysed by Locke (1689)

electrician tries to work out the course of electric wiring in a wall. In all these cases nothing physical is happening but in each case the processes have goals that are intentional and are not the result of unconscious thought. If, for example, a person succeeds in remembering where his keys are then the action has been successful but not otherwise.

Given that mental actions resemble ordinary physical actions, they share the characteristics of physical actions. Since they are intentional then they are caused by reasons comprising beliefs and desires (using Davidson's (1963) analysis) and, given our use of the expected utility model, these can be measured using probabilities and utilities respectively. The mental action that we are interested in is that of the mind fitting a particular representation to the situation in which the agent finds themselves (Proust 2009) and we will use the phrase "mental action" to refer to this type of action from now on.

It can be seen, therefore, that representations can be modelled within game trees where there are two types of actions. The set of actions in a game $\mathbb{G}$ therefore will consist of physical actions $b_m \in \mathbf{b}$ and mental actions $T_j \in \mathbf{T}$. Mental actions in these games will denote the change of representation for all outcomes subsequent to T_j. Throughout the rest of this paper we will assume that a player chooses T_j, triggering representation $A(\mathbf{x})$ and chooses actions which lead to an outcome s_k with associated predicate q_k. The utility which we will analyse will be that both $A(\mathbf{x})$ and q_k occur i.e. the conjunction of the two predicates: $(A(\mathbf{x}) \wedge q_k)$. This indicates that the effects of both the representation and the outcome are being considered within the utility function. However we will also impose the condition that $A(\mathbf{x})$ and q_k are additively separable to reflect the idea that they are independent of each other in that $A(\mathbf{x})$ is created by the player while q_k is a reflection of the possibilities logically available to the player if s_k is chosen.

It should be noted that how mental actions enter into games may vary from game to game. Conventional games have no explicit mental actions since they do not change representations- the mental action actually occurs before the game starts when a representation is chosen to fix the utilities for that game. Other games may involve a mental action and physical action happening simultaneously- such as when one goes through the ritual for converting to another religion. However, quite often, mental and physical actions may be separate in time within the game.

Following Rubinstein, such games will involve replicating the individuals' subjective perception of a situation. In some games, different mental actions will result in different

choice sets for physical actions whereas in others the choice set is effectively replicated for different mental actions- just the utilities of the outcomes change even though, physically, nothing else has. In other games, most of the game tree will operate under one representation but a mental action will give the opportunity to change to another representation in a particular part of the tree.

Finally, the use of representations automatically implies the use of *extensive form games* rather than normal form games. This is because there is a natural time difference between the choosing of a representation and the subsequent choosing of an option from the basic game. This is necessary because the choice of a representation informs the interpretation of the situation that results in preferences over options in the choice set. Representations must be chosen first so that options can be chosen with coherent preferences[12]. It follows from this that we will have a preference for using extensive form refinements when looking at equilibrium in these games.

To some, this would seem to be a recipe for disaster. Surely a change in representations halfway through a game will result in inconsistency? If one changes representations then this results in changes to utilities so that utilities as perceived in one part of the game tree will not be valued in the same way in other parts. As Rubinstein has pointed out (Rubinstein 1991), if one wants to model the subjective perceptions of individuals in games then there will always be some inconsistencies. However, this still leaves the problem of how we can analyse such games. The next section will discuss this point.

6) Conditional Utilities and Representations

As constructed, the framework given here has problems that need to be solved. Assuming that representations can change partway through a game, how can a player come to an overall assessment as to which strategies are the best to play? In essence, a player will have to put himself in the position of actually having that representation and seeing what his utilities are *given* that representation. This suggests that we will have to analyse this situation in terms of conditional utilities. Furthermore, we will specify two types of such utilities:

[12] In the examples later on in this paper, it will sometimes be assumed that agents already have a fixed representation or that a move by one's opponent rigidly determines one's own representation. This does not detract from the point made here.

subjunctive conditional utilities and *indicative* conditional utilities. Both are necessary to analyse the problem of inconsistency.

Another issue is how exactly one should judge the representations themselves. Are representations close to reality? Do they have high values for accuracy and objectivity? Do they reflect the importance attached to various values by the player? These issues do not necessarily link in with the utility of individual outcomes but could be common to all outcomes following from the relevant mental action.

The theory of conditional utilities we will use here is based (with some changes) on that of Weirich (1980). In his view, conditional utilities relate to the utility, given a condition, of an outcome supposing that that outcome occurs. It is not seen as the utility of a conditional. Weirich holds that there are different ways of supposing the conditions but that the two most fruitful are suppositions for indicative conditionals and for subjunctive conditionals. In the former case one looks at what happens if a condition *does* hold compared with the latter where one looks at what happens if a condition *were* to hold.

The difference may at first sight seem slight but the effects are profound in the game utilities presented here. For indicative conditional utilities, one's preferences for an outcome are determined given that one *has already accepted* the impact of the representation on the outcome. For subjunctive conditional utilities, one imagines what it *would be like* to have that representation. This means that there could be a considerable difference in the utilities assigned to a particular outcome, depending on whether one is assessing the game tree before playing it or whether one has actually played it.

In Weirich's notation, if we have a consequent L and condition N then the condition has a subjunctive conditional utility denoted by $U^*(L \wedge N)$, where both L and N are supposed at the same time. The corresponding indicative conditional utility is denoted by U(L/N) where L is supposed given that N actually holds. The two types of utility are assumed to be equal when $U^*(L \wedge N)=U(L/N)$. It should be noted that the indicative utility U(L/N) is similar to the conditional utility defined in Keeney and Raiffa (1976)for attributes and it will be defined in that sense here. Weirich assumes that the subjunctive conditional utility is the "normal" way of assessing utilities in decision problems and games given the use of subjunctive reasoning throughout a game (c.f. Binmore 1987).

Another aspect of conditional utilities that needs to be discussed is the nature of the condition itself. Using the notation introduced in the previous section, simply assuming that the indicative conditional utility is formulated as $U(A(\mathbf{x})\wedge q_k/A(\mathbf{x}))$ ignores whether $A(\mathbf{x})$ imposes any restrictions on q_k (Weirich 1980). The mere choice of T_j and the taking on of $A(\mathbf{x})$ as one's representation does not say anything about the causal link between the representation and the subsequent outcome. It follows that it would make more sense to make the restriction on q_k more explicit. Hence a better formulation would be $U(A(\mathbf{x})\wedge q_k/A(\mathbf{x})\rightarrow q_k)$ where "$\rightarrow$" represents a conditional that imposes $A(\mathbf{x})$ as a restriction on q_k[13]. Three assumptions made about the conditional $A(\mathbf{x})\rightarrow q_k$ are that, for all k where s_k is a successor to T_j, it is preferentially independent of its complement, for k=i the restriction $A(\mathbf{x})\rightarrow q_i$ is the only one that is relevant to the outcome s_i and $A(\mathbf{x})$ conditioned on a condition other than $A(\mathbf{x})\rightarrow q_i$ is not relevant to the utility of s_i. This holds because we do not wish to add the complication of analysing what happens when the content of one outcome enters into the utility of another, such as when under the influence of regret (e.g. Loomes & Sugden 1982). In addition we do not want to consider any influence external to the outcomes apart from that of $A(\mathbf{x})\rightarrow q_k$.

Finally, we need to think about that part of $A(\mathbf{x})$ that does not imply q_k for any k. For this part, we will assume that the subjunctive conditional utility is always equal to the indicative conditional utility. This is because this is the part of $A(\mathbf{x})$ that is *unavoidable* for the player and hence includes elements of $A(\mathbf{x})$ that are common to all options and outcomes succeeding T_j. This would include part of the utility of $A(\mathbf{x})$ that is involved in judging whether it is an appropriate representation to hold or not. Assuming that the subjunctive and indicative conditional utilities are equal allows us to assume that this part of the utility of $A(\mathbf{x})$ can be compared across the whole game[14].

[13] The conditional link denoted by "→" is not intended to be a material conditional but rather a conditional between the representation and the outcome in the sense that A(x) restricts qi. The material conditional is controversial as a good proxy for the common notion of the conditional relation, especially with regard to the truth functional idea that the material conditional is always true if the antecedent is false (see Sanford 1989). In this paper two assumptions are made that seem reasonable: modus ponens holds so that Z, Z→C C; also it is the case that if a conditional is false then the antecedent is true and the consequent is false so that ¬(Z→C) Z^¬C.

[14] This assumption allows us to ignore problems related to infinite regresses of representations used to check whether a particular representation is correct. It is simply assumed that the procedures used to assess the utility of A(**x**) are the same as for assessing other utilities in the game.

This allows us to prove the following proposition that decomposes the utility function of $U(A(\boldsymbol{x})\wedge q_i)$:

Proposition 1: If the subjunctive and indicative conditional utilities are equivalent then the utility of the outcome s_i which is subsequent to the mental action T_j can be expressed as:

$$U(A(\boldsymbol{x}) \wedge q_i) = U\big(A(\boldsymbol{x})\big|\forall k\colon \neg(A(\boldsymbol{x}) \rightarrow q_k)\big) + U(A(\boldsymbol{x})|A(\boldsymbol{x}) \rightarrow q_i) + U(q_i|A(\boldsymbol{x}) \rightarrow q_i)$$

when expressed as indicative conditional utilities or:

$$U(A(\boldsymbol{x}) \wedge q_i) = U^*(\forall k(A(\boldsymbol{x}) \wedge \neg q_k)) + U^*(q_i)$$

when expressed as subjunctive conditional utilities.

(Proof in appendix)

Here U* represents the utility function for subjunctive utility functions whereas U represents the utility function for the indicative utility functions.

Taking the indicative conditional utilities first, the first term on the right hand side is the indicative form of the utility of the representation when we ignore the effect of it on options in the game. This would include the player's judgement of whether the representation A(x) fits with the situation being assessed. The first term on the right hand side of the second equation is the same idea expressed in terms of subjunctive utilities. These represent the unavoidable utility component as explained above and would include an element that represents the "fit" of the representation to the outside world.

The second term on the right hand side of the first equation represents the utility of the representation that emerges from consideration of playing outcome s_i. This is where the importance of the representation is entwined with the playing of a particular option. In other words, the playing of a particular option justifies the holding of a particular representation while the playing of another option (i.e. one that undermines the whole purpose of the representation) would not do so. This could also be seen as part of the option-specific "fit" to a situation.

The final term on the right hand side of the first equation is the utility of the outcome s_i. It should be noted that this assumes that the utility is fully determined by the representation and there is no separate influence from elsewhere. This term, combined with the second term is equivalent to the second term in the second equation which is the subjunctive evaluation of

the outcome. This is the *avoidable* part of the utility component which is so-called because it can be avoided by choosing another option.

This proposition suggests that we can model multirepresentational games as a two stage decision process within an extensive form game with the utilities of the representation and the chosen outcome summed at the end. Once the subjunctive and indicative conditional utilities are equal to each other the game can be treated as an ordinary game, using ordinary game- theoretic solution concepts.

One interesting aspect of this is that, once the individual is in equilibrium, he is in equilibrium in both physical actions *and in representations*. The individual is not free to simply change his interpretation of the world at whim because to do so would be to decrease his expected utility. Naturally, the individual could acquire another, third, representation but this would change the nature of the game by adding another mental action. Nevertheless, this new game could be solved in a similar way and one of the representations would become fixed in the new equilibrium. Allowing for an agent to interpret a situation within a game does not lead to theoretical anarchy but instead explains the stability of these interpretations in the form of equilibrium representations.

One question that can be asked is under what conditions an individual's subjunctive conditional utility will converge on their indicative conditional utility. This is a complex question that will have several different answers depending on the formulation and number of representations in the game. We will focus on the total avoidable utility in the game i.e. $U(A(\mathbf{x})\wedge q_i/A(\mathbf{x})\rightarrow q_i)$ and the corresponding subjunctive utility. For the purposes of this paper we will simplify issues so that there are just two representations $A(\mathbf{x})$ and $B(\mathbf{y})$ and that these are the only two possible representations in a player's mind. $B(\mathbf{y})$ is constructed from the set $D(\mathbf{y})$ mentioned above by deleting the set $C \subset D(\mathbf{y})$ where C comprises those complex predicates of $D(\mathbf{y})$ that contradict complex predicates in $A(\mathbf{x})$.

This allows us to formulate the notion of a subjunctive conditional within our notation. If a player is coming to a representation $A(\mathbf{x})$ that is not currently active then they must have another active representation that we will assume is $D(\mathbf{y})$. According to Papineau (2012 p. 116), the new information represented by $A(\mathbf{x})$ is assessed by taking away all predicates that contradict $A(\mathbf{x})$ thus creating $B(\mathbf{y})$ and then adding A(x) to our stock of knowledge. This creates a new predicate: $Q(T_j,\mathbf{x},\mathbf{y})=A(\mathbf{x})\wedge B(\mathbf{y})$. $Q(T_j,\mathbf{x},\mathbf{y})$ therefore relates to

the perception of the player when assessing the game rather than actually playing the game. It follows that the subjunctive conditional utility is denoted by $U^*(A(\mathbf{x}) \wedge q_i \wedge (Q(T_j,\mathbf{x},\mathbf{y}) \rightarrow q_i))$ where $(Q(T_j,\mathbf{x},\mathbf{y}) \rightarrow q_i)$ is the adjusted condition for the subjunctive conditional[15].

Define the constant vectors $\mathbf{x_0}$, $\mathbf{x_1}$ as instances of $\mathbf{x}$ such that:

$$\forall x_{in}, x_{im}((x_{in} \in \boldsymbol{x_o}) \wedge (x_{im} \notin \boldsymbol{x_0})) \leftrightarrow (x_{in} < x_{im})$$

and

$$\forall x_{in}, x_{im}((x_{in} \in \boldsymbol{x_1}) \wedge (x_{im} \notin \boldsymbol{x_1})) \leftrightarrow (x_{in} > x_{im})$$

These are the minimal and maximal exemplars of $\mathbf{x}$, and we can similarly define exemplars of $\mathbf{y}$, labelled $\mathbf{y_0}$ and $\mathbf{y_1}$.

This allows us to prove the following proposition:

Proposition 2: Assuming mutual utility independence between the attribute vectors $\mathbf{x}$ and $\mathbf{y}$ and that the prospects:

$(A(\mathbf{x}) \wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0}) \rightarrow q_i))$ with probability π; $(A(\mathbf{x}) \wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0}) \rightarrow q_i))$ with probability $(1-\pi)$ and

$(A(\mathbf{x}) \wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_1}) \rightarrow q_i))$ with probability π; $(A(\mathbf{x}) \wedge q_i \wedge Q(T_j,\mathbf{x_0},\mathbf{y_1}) \rightarrow q_i))$ with probability $(1-\pi)$

are indifferent for all probabilities π then the player will hold the correct representation $A(\mathbf{x})$ and the indicative conditional utility $U(A(\mathbf{x}) \wedge q_i / A(\mathbf{x}) \rightarrow q_i)$ will be equal to the subjunctive utility $U^*(q_i)$.

(proof in appendix)

The proposition above shows the link between subjunctive and indicative utilities, demonstrating how they can be achieved in the particular circumstance where there are two representations. However, it may be argued that there may be many circumstances where equality between the two types of conditional utility do not hold. It may be legitimately asked whether this matters. One answer is that it won't matter when the indicative conditional

[15] It is possible, given similar preferential independence assumptions, to replicate the analysis of proposition 1 for Q(*).

utility $U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i)$ is monotonic with the subjunctive conditional utility $U(A(\mathbf{x})\wedge q_i \wedge (Q(T_i,\mathbf{x},\mathbf{y})\rightarrow q_i))$. This implies that the former has the same preference ordering as the latter so the difference between the two utilities won't affect which option is chosen.

The conditions for this can be seen in the following proposition:

Proposition 3: If two representations defined by the attribute vectors **x** and **y** are mutually utility independent then **x** and **y** are complements if and only if $U^*(A(\mathbf{x})\wedge q_i \wedge(Q(T_i,\mathbf{x},\mathbf{y})\rightarrow q_i))$ is monotonically increasing in $U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i)$.

(Proof in Appendix)

If it is possible to observe the attribute vectors and their relationship to each other then it follows that one can predict whether or not it matters if the subjunctive conditional utility is aligned with the indicative conditional utility.

7) Discussion

One immediate objection that springs to mind is that this seems to put forward a bizarre view of how representations are selected. If one *chooses* which representation is most appropriate then why not choose one that suits oneself? Why go to the trouble of determining the accuracy of a representation and then operating within the (possibly inconvenient) confines of this representation when one could maximise one's utility by imagining a representation that best suits one's own tastes?

The primary response to this is to point out the implicit assumption made by this objection. It assumes that people are self-interested in every possible way so that they are willing to warp their view of reality to fit in with their self-interest. However, the model in this paper explicitly does not assume universal self-interest. This means that we are able to assume, with Williams (2002), that accuracy is seen as important i.e. it has value and that this value is incorporated in the utility function. Self-interest and accuracy do sometimes go together but quite often they diverge- one's plans are often blocked by "inconvenient facts". Furthermore, the values assigned to representations need not be self-interested. Values indicate importance and this need not imply that something is only important to oneself. However, on the other hand, we would not want to exclude the role of wishful thinking so in

this model it is possible that people *may* solely follow their own interests. Wishful thinking, for example, occurs when accuracy is *not* seen as being valuable and when one's personal desires overwhelm the facts on the ground.

In spite of the above considerations, it is still a valuable question to ask just how accurate one would expect humans to be. There is an evolutionary argument to expect respect for accuracy to be widespread. Any set of individuals who do not value accuracy are highly likely to find themselves in dangerous situations where they cannot evaluate properly what is happening. Such people are liable to be killed off, leaving those who value accuracy with a high likelihood of surviving. This does not mean that wishful thinking is driven to extinction- in non-lethal situations and, in situations where one's decisions do not necessarily have an impact on oneself, it would survive but the habit of valuing accuracy would give an advantage. There is room for wishful thinking and self-deception in human society but one would not expect it to be pervasive.

Another, similar, objection relates to the fact that a decision tree approach would result in a person aggregating their expected utilities in their terminal outcomes in a game tree. It could be argued that this is a very odd way of doing things because the utilities resulting from the choice of representation are combined with the utilities assigned to outcomes as a result of adopting the representation. There is no privileged status for the representational utilities so one could have a representation chosen by the latter utilities overriding the former.

Again, this need not be a problem. The "fit" of the representation to the external world is not trivial in utility terms. To take a fanciful example: suppose a person is working all day earning money as a typist. He is faced with a decision as to whether carry on working in the job or not. He has two representations: one is realistic and results in a decision to carry on working as this is the only way in which he can support himself. The second representation is fantastical in that it assumes a fairy is going to endow the typist with a lot of money so that he doesn't need to work ever again. While the utility from the money acquired in the second representation may be enormous, it would be swamped by the expected utility

gained from the comparative "fit" of the first representation and the typist's valuation of accuracy[16].

There may, indeed, be cases where, for example, the overall fit of two competing representations is so low that the utilities of the outcomes are decisive. It may be the case that this decisive advantage is the result of the importance of the representation as expressed through a particular option being played i.e. the option-specific "fit". In this case there is no problem since representations' utilities still have a decisive influence over the choice between representations. If, however, these are also low then one has little real evidence for the reality of the two representations or one values accuracy at such a low level that the outcomes are as good a way of judging which representation to choose as anything else.

Another case, which will occur in the examples, is if the fit of the competing representations are both comparatively high, in both overall and option-specific terms, so that there is little to choose between them purely in terms of the fit. However, I would argue that this is not a problem; if both of one's representations fit the world well then one *would* make a choice by looking at which is the most internally attractive in terms of beliefs and desires.

Another obvious objection to the representations framework is that when this is conceptualised as a game, one's opponent cannot see which move has been made as it occurs inside the opponent's mind. However, there are three possible responses to this objection. The first response is that hidden moves are easily modelled in game theory through the use of non-singular information sets. The fact that an agent does not know how another agent has moved is easily modelled in an extensive form game. It should be noted, however, that it is assumed that all *possible* representations chosen by the other agent are known even if one does not know the *specific* representation that was chosen.

The second response is that, for various reasons, the representation chosen will be "obvious". To take an extreme example, a person is unlikely to have a representation that will result in him wanting to drinking poison. In other examples, "obviousness" may be culturally specific. In China, for example, it is generally assumed that Chinese people are able to use chopsticks. In other cases, people may consistently have a psychological bias against certain

[16] This does not preclude people from choosing the other way, although one might categorise those who do so as "fantasists".

things. For example, people tend to find furry animals to be more "cute" than (say) reptiles and may be more likely to pet them.

The third response derives from Frank's (1988) work on emotions. According to Frank, emotions act as signalling devices because they are difficult (although not impossible) to fake. If a representation causes emotional arousal in an agent then the other agent may be able to detect that emotion and distinguish which representation their opponent has chosen. Emotion is costly for the person involved and, being difficult to fake, acts as a commitment device that makes the signal reliable. One would expect this method to be used quite often in personal interactions.

Another issue relating to representations as game moves is that of mixed strategies. We have assumed that individuals are rational in the sense that they do not accommodate contradictions. As a result of this, an agent would not be willing to hold two contradictory representations at the same time. However, with the traditional interpretation of mixed strategies this is precisely what does happen. Given equal expected utilities for representations then one will randomise over the two representations even if they have elements in them that contradict each other.

However, this need not be a major problem in games and mixed equilibria can still be used. This is because there are alternative interpretations of the mixed equilibrium concept that can be used instead. One alternative interpretation was put forward by Aumann (1987) where he interpreted a mixed strategies equilibrium as being a state where players' subjective probabilities were over their opponents' chances of playing one pure strategy (or representation) over another rather than randomising one's own strategy. Another alternative is to follow evolutionary game theory where a mixed strategy is interpreted as the proportions of a population who are playing a given pure strategy (or representation).

8) Representations, corruption and consent

Having outlined how representations can be modelled in a game, we will need to give examples of how this may be done in more concrete situations. Fortunately, as has been described, representations can be modelled in much the same way as actions in a game, if one allows for the facts that they are largely hidden from opponents, that one cannot use the traditional view of mixed strategies, that games are interpreted as the agent's perception of a

given situation and if it is assumed that subjunctive and indicative conditional utilities are equivalent.

Furthermore, we need to link our discussion of representations to the corruption of goods. We have rejected the idea that this can be modelled as a type based analysis as suggested by Benabou and Tirole (2006). This type of analysis ignores the role of human decision making in adopting representations and has no explanation as to why players may change their minds. Any corruption of a good, therefore, is better modelled as a *decision* by an agent that the good does not have the same value as it had before.

In this section, we will model corruption of goods using psychological game theory (Geanakoplos, Pearce & Stacchetti 1989) where fundamental deviations from self-interest[17] are modelled on an individual's beliefs about their opponent's actions. It follows that differences in utility resulting from differences in representation will be determined by one's own beliefs in one's opponent's actions.

In discussing the modelling of representations we will model three examples derived from the discussion in section 2. The examples given will be simple in form but they will highlight some of the issues mentioned previously in this paper. The first case is that of consent (Peter 2004). The example given was of the NHS in the United Kingdom and the change from public to private providers of healthcare. The second case is that of the Swiss nuclear waste repository where individuals approve of having a nuclear waste repository without incentives but reject it when they have it. The final case relates to bribery and whether a person will accept a bribe or not.

The first case is modelled as follows: the government decides on what sort of service (e.g. private or public) is provided for healthcare. There are three possible options (labelled a,b- two treatments and e- an exit option) that the second player, an individual consuming the country's health service, can choose between these options.

[17] Self-interest here is interpreted fairly broadly so that government action, for example, is seen as "self-interested" if the government is following national interests that do not necessarily align with other players in the game.

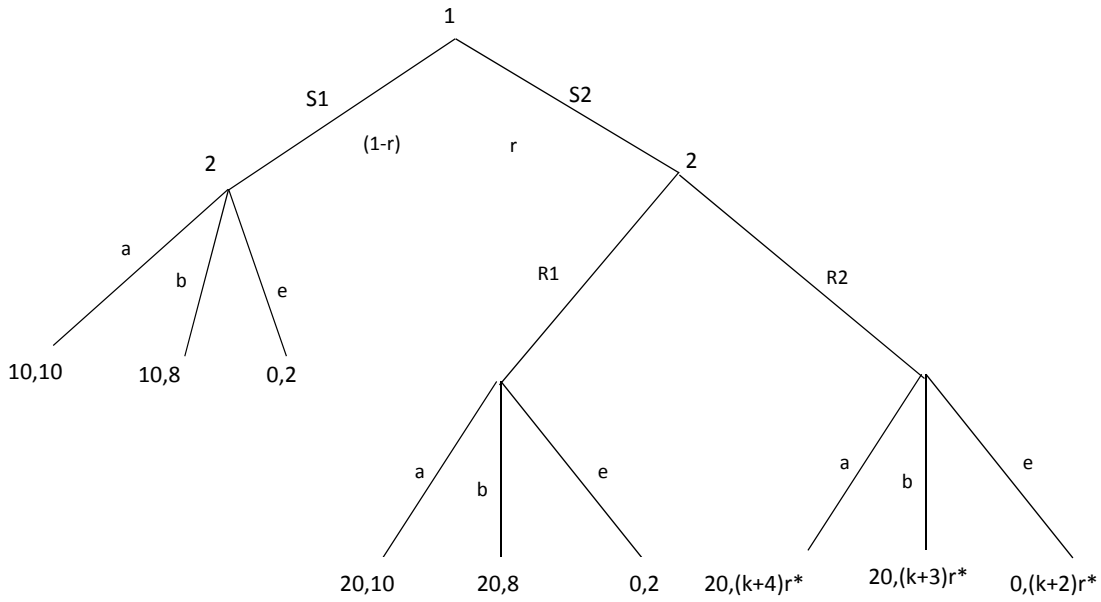

Figure 1: Healthcare game

In the diagram above we can see that S1 and S2 represent public and private provision of the service respectively. S1 is the "status quo" option in that it is assumed that the service has been public for a long time and so the second player's reaction to this style of service has settled down on one active representation, which we will denote R1. From the viewpoint of player 2 there is a (1-r) probability of S1 being chosen and an r probability of S2 being chosen. There are two possible representations held by player 2: R1 and R2. R1 is the active representation that is held when public service is continued while R2 is a possible alternative representation that holds when private service is imposed. When private service is imposed, the second player has to make a choice between the two representations and then, conditional on each representation, a choice between the treatments provided.

It is assumed that the government approves of the change in service and so prefers people to have private service rather than public service but doesn't like it when people exit from the service altogether. The individual's utilities are controlled by the representation they have of the situation. We will assume that individuals always prefer treatment a to b and both of these to the exit option. In addition, the avoidable utility does not vary according to the importance of a representation when an option is played.

Under representation R2, player 2 has utility payoffs depending on the player's expectation of r, r*. In each action, after playing R2, part of the payoff is composed of kr* with k>0 being the unavoidable utility that includes the fit of R2 to the situation. This means

that if kr* is high then the representation R2 is deemed to be a good "fit" to the situation. The "fit" of R1 is normalised to zero in this model. r* is seen as part of the "fit" because the change from R1 to R2 is triggered by player 1's choice of S2. The choice of S2 causes player 2 to reassess his values and beliefs in response. The avoidable utility varies between the actions with action a having a payoff of 4r*, action b having a payoff of 3r* and action e (the exit option) having a payoff of 2r*. This part of the payoff (apart from e) can be seen as a *perceived* future decline in the quality of healthcare.

It follows that r*=r=1 is the only possible probability assignment for any psychological subgame perfect equilibrium. This can be seen by realising that action a dominates the actions b and e for player 2 throughout the game. The point r*=r=0 is not consistent since it produces an equilibrium {S2,R1,a} which cannot occur if r=0 . The mixed equilibrium would be where r*=10/(k+4). This does not exist if k<6 and is the same as r*=r=1 if k=6. However, if k>6 then this creates a mixed equilibrium in a situation where player 1 will always play S2 (because action a is always chosen and actions a following S2 always have a payoff of 20 for player 1 while the action a following S1 has a payoff of 10). This means that the mixed equilibrium is also not consistent.

It follows that one can analyse the game purely in terms of the unavoidable utility k. If k > 6 then a play of {S2,R2,a} becomes the psychological subgame perfect equilibrium of the game. If k < 6 then {S2,R1,a} is subgame perfect while k=6 involves player 2 being indifferent between the two. In both cases, whether the service remains in the public sector or not, the type of treatment chosen remains the same and, from the point of view of physical actions, there is no difference.

However, this should not be taken as approval of the system under which treatments are provided. If k > 6 and R2 is played then the avoidable utility attached to the *treatments* has actually gone down i.e. the change in system for providing the treatments has "corrupted" them. The reason for choosing R2 is because of its perceived fit to the situation (with unavoidable utility of kr*), not because the treatments available look better. However, the individual has not chosen the exit option since it still looks like a poor alternative within the game. This, as Peter (2004) points out, does not imply consent. Indeed, if a survey was taken then considerable dissent would probably be expressed. This means that "voice" is a necessary part of institutional design and we cannot rely on exit options alone to indicate dissent.

It should also be noted that from the point of view of the outside observer, this game would seem to be very simple. If we ignore the change in supplier then the same treatments are being produced under public or private provision. One may see this simply as a straight choice between two treatments and an exit option. We can also include the government's choice of two types of service but this does not allow us to see the thought processes inside an individual's head. This can only be done with representations.

Case 2 revolves around the situation of the Swiss state wanting to place a nuclear waste repository in the village of Wolfenschiessen. In the game given in figure 2, the investigators either offer (O) or do not offer (DO) money for the villagers to accept the repository. The status quo situation is that the villagers are not offered money. In this case they are assumed to subscribe to their original representation (labelled R1) and immediately decide whether to accept (A) or decline (D) the waste depository. If the villagers are offered money then they make a choice between two new representations: R2 or R3. R2 is a representation where player 2 has monetary payments as motivation to accept the depository, while R3 is a representation where monetary payments, while motivating to a certain extent, also cause a reaction against them. When they have chosen their representation, they accept or decline the repository.

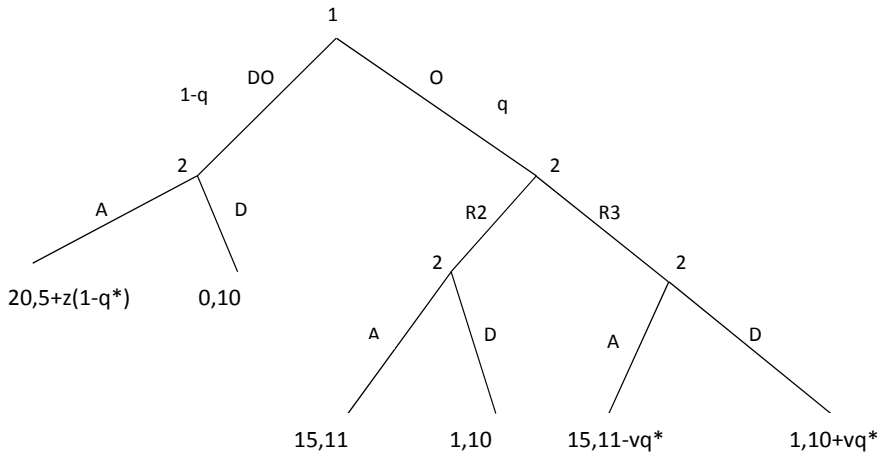

Figure 2: Nuclear Waste Game

It should be noted here that the unavoidable utility derived from the "fit" of the representations is not assumed to have much impact in this case. This would imply that the

representations are seen as equally plausible as a good fit for the situation and so that the variable k used in the last example would be normalised to zero for all three representations. The main effect of the different representations is to change the balance of utility payoffs i.e. the avoidable utility between A and D. This change can be seen as a result of the perceived importance of the representations when A or D are played as well as the utilities induced by the representations.

It is assumed here that the offering of money will result in a change in representation whatever happens. For player 2, there is a difference between what happens to the payoffs if the repository is accepted. Under R1, player 2 will gain an avoidable utility of $5+z(1-q^*)$ reflecting a situation where the government is relying on people's goodwill to gain acceptance. The decision not to pay is reflected in the payoffs. By contrast, under R2, the initial baseline utility of 5 is supplemented by 6 to give an overall utility of 11. Under R3, there is yet another change whereby for player 2, the additional utility from the payoff of 6 when the depository is accepted is offset by a utility loss of vq^*. This utility loss is transformed into a utility gain when the repository is rejected. Both v and z are constants where z is the initial level of social goodwill, while v represents a combination of outrage at the money payment and a rejection of the social good. It should be noted here that, when player 1 decides to pay the residents, the payment represents a utility cost of 6 but this is partly made up by a preference for using monetary incentives by a factor of 1 across the board.

There are three psychological subgame equilibria possible in this game, depending on the values of the parameters. There are no mixed equilibria in this game. This is because, for player 1, if DO is chosen then player 2 can choose either A or D. If A is selected then, for player 1, the payoff of 20 dominates the payoffs from any choices made by player 2 after O is chosen. If D is selected after DO then, for player 1, the payoff of 0 is dominated by the payoffs from any choices made by player 2 after O is chosen.

The pure strategies psychological subgame perfect equilibria depend on the values of z and v. If $q^*=q=0$ and $z>5$ then the government doesn't offer money and the individual accepts the depository. If $q^*=q=1$ then there are two possible outcomes depending on the value of v. If $v>1$ then there is an equilibrium with strategy profile {O, R3, D}. If $v<1$ then there is an equilibrium with strategy profile {O, R2, A}. It follows that this model demonstrates how the various outcomes can emerge with different representations.

In a similar manner to case 1, to the outside observer, the decision looks very simple- the villagers simply have to choose between accepting the waste depository and rejecting it but the use of representations shows that there is actually a rich underlying decision process. Also, in the same way as with case 1, the representations are visible to the villager making the decision but is not visible to the state so that the state cannot "see" whether one representation is chosen or another.

The final case focuses on the issue of bribery as discussed by Grant (2012). We will assume in this case that there is a judge (player 1) who has been offered a bribe in order to carry out a particular judicial decision. We will model this as a choice by a particular judge as to whether they should see the bribe as a commercial transaction (and choose R1) or as a moral issue (in which case they will follow R2). It should be noted that we are not assuming absolute morality here. Given enough money the judge will be won over or show weakness of will[18]. We will also assume, as with the second example, that the unavoidable utility (including the overall "fit" of the representations) does not affect the outcome in this example. However, the choice of mental action will be determined by the importance of the representations when different physical actions are chosen.

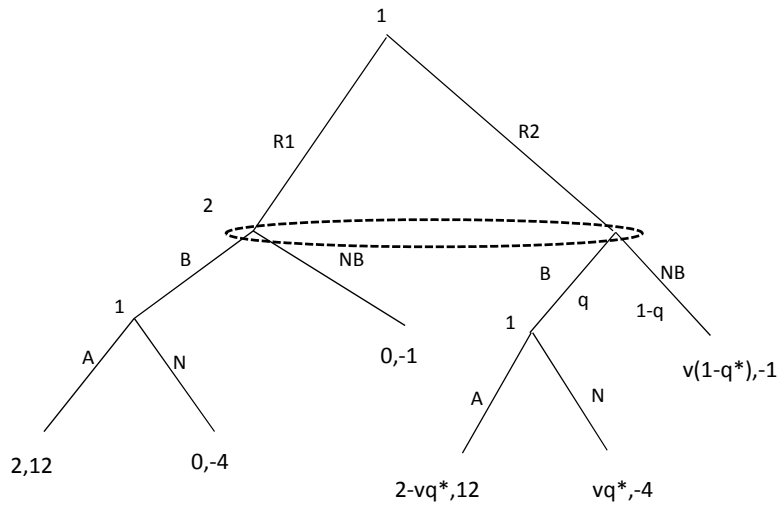

Figure 3: Bribery Game

[18] One way to demonstrate an absolute aversion to corruption would be to have utility set at infinity.

The choice between R1 and R2 is hidden from the briber (player 2) and so there is an information set covering the starting nodes of the next decision. This is the decision as to whether to bribe the judge (choosing B) or not (choosing NB). p is the probability of assuming the representation R1. v is a constant representing a norm against bribery. If it is decided not to bribe the judge then the briber gets a payoff of -1, while the judge gets a payoff of 0 if they are corrupt (i.e. choosing R1) but v(1-q*) if not corrupt (i.e. choosing R2). q* is the expected value of q- the probability that the briber chooses B. If a decision is made to bribe then the judge has to decide whether to accept the bribe (choosing A) or not (choosing N). If the judge accepts the bribe then he gets a payoff of 2.If a bribe is not accepted then, given the possibility of getting caught, the briber gets -4 while the judge gets zero. Finally, under representation R2 there is a utility value vq* that reduces the utility value of accepting a bribe if offered and boosts his utility if it is refused.

Obviously this model is complex and will have different equilibria depending on the values of v. We will focus on the pure strategies equilibria. There are two possible situations depending on whether $v \leq 1$, where only one psychological sequential equilibrium is possible, and $v>1$ where three psychological sequential equilibria are possible. The former can be seen by realising that B will be chosen by the briber if the payoff is greater than that for NB. On the left hand side playing B will always result in A being chosen since 2>0 for the judge. This means that playing B on the left hand side always results in a payoff of 12 for the briber. On the right hand side the payoff for B depends on whether 2-vq*is greater or less than vq*. If 2-vq* is greater then the payoff for the briber is 12 so playing B dominates playing NB for the briber on both sides. Therefore there is one psychological sequential equilibrium where q=q*=1 and the strategy profile {R1,B,A} is played.

Where v>1 there are three psychological equilibria depending on the value of p. If $p<3/16$ then NB will be selected on both sides so q=q*=0 and, since v(1-q*)>0, a strategy profile of {R2, NB} will be played. If $p> 3/16$ then B is generally selected so q=q*=1. The choice of player 1 depends on the value of v. If $1<v<2$ then there is a strategy profile of (R1,B,A}. If v>2 then the strategy profile chosen will be {R2, B,N}.

It follows that this game allows for a variety of different moral stances to be taken by the judge in different situations. It can be seen that in a situation where morality is given low weight (i.e. $v \leq 1$) then the judge will become commerce minded, be offered the bribe and accept it. If morality is given a higher weight (i.e. v>1) then the situation is not so clear. Only

if v>2 will the judge refuse the bribe if offered. At lower levels (1<v<2) the judge may accept the bribe while being commercially minded but, interestingly, the prospect that the judge may be morally- minded may dissuade the briber from bribing the judge at all.

Conclusion

Recently, Economics as a discipline has come under attack from various quarters. Some of these attacks are unjustified and come from a lack of understanding of the economics discipline. However, some of these attacks expose little-discussed weaknesses in the armoury of economic theory. One such weakness, as exposed by Michael Sandel and Ruth Grant, is the problem of changes in attitude or, more specifically, the corruption of goods. The main argument of this paper is that this problem can be tackled fairly easily as long as one is willing to extend one's view of modern game theory to include the idea of representations.

This change has huge advantages if it can be done successfully. First of all, it provides a better fit with other disciplines under the cognitive science umbrella. The use of representations as a modelling tool is normal usage within cognitive psychology, artificial intelligence and philosophy of the mind. The failure to acknowledge them within economics is peculiar and means that many conversations across disciplines are effectively closed off to economists.

Secondly, the use of representations provides a more natural modelling technique for certain situations than the ones currently available. The use of Harsanyi's type theory has had a profound and positive effect on economic theory, expanding the ability of economists to explain the economy. However, its focus on fixed types of individuals, however ingeniously applied, makes a very awkward fit for situations where individuals are obviously moving from one set of attitudes to another i.e. their types are *not* fixed. The method outlined here tries to fill in the gap in the most parsimonious way possible. Given that consistency conditions hold, one can model varying attitudes as a choice in representations within a game theoretic context.

To a certain extent this also solves some of the problems of *ad hocery* that plague such behavioural models. While the applications outlined in the previous section do involve parameters to operationalise each representation, these are less arbitrary than in the Benabou

and Tirole model. These parameters take effect under certain conditions- when their representations are chosen- and their range of operation is defined by them. When they are triggered and in how much depth also depends on the psychological equilibrium that the game converges on. Also,the use of representations also allows the wide range of research in psychology on this topic to be exploited (Agoustinos 1995).

Finally, the model allows a solution to the problem of the corruption of goods. Goods become corrupt when attitudes towards them change. In the examples put forward by Grant and Sandel, these usually involve situations where non-monetary transactions are converted into monetary transactions. In such a case the monetary reward for undertaking an action does not enhance the individuals' desire to carry out that action but may undermine it. The desire to carry out the action, therefore, has been corrupted by the money. This change in desire is modelled as a change in representation resulting from a deliberate choice on the part of the subject on perceiving that the situation has changed.

The model does have weaknesses and these are an obvious focus for future research. One obvious one is that there is no guarantee that the indicative and subjunctive utilities will coincide or even be monotonic with each other. This likelihood will become greater as the number of representations increases from the binary system investigated here. It will be necessary to find out how humans in practice manage to maintain consistency or, failing that, how they deal with the inconsistency if they do not coincide.

The model can obviously be extended in multiple ways. One way would be to use evolutionary game theory to see how populations of players can converge on one representation or another. This would imply that all players in a game would have the opportunity of changing representations. The implications of this are potentially of great interest. One could study the spread of values, cultural mores and habits across a population and gain insight into many social processes.

Another area of interest is in the thought processes involved in wishful thinking. There are many examples where wishful thinking seems to rule supreme in areas such as alternative medicine or conspiracy theories. How do populations come to converge on truthful representations rather than wishful scenarios? Other applications arise when considering expectations of future investment, negotiations in the bargaining process etc.

Bibliography

Agoustinos M. (1995) “Social Cognition: An Integrated Introduction” Sage London

Aumann R. (1987) “Correlated Equilibrium as an Expression of Bayesian Rationality” Econometrica 55 1 1-18

Benabou R. & Tirole J. (2006) “Incentives and Prosocial Behavior” American Economic Review 96 5 p. 1652-1678

Besley T. (2013) “What’s the Good of the Market? An Essay on Michael Sandel’s *What Money Can’t Buy*” Journal of Economic Literature 51 2 478-495

K. Binmore, “Modeling Rational Players I and II”, *Economics and Philosophy*,3, p.179-214 and 4, 1988, pp.9–55.

Bowles S. & Polania Reyes S. (2012) “Economic Incentives and Social Preferences: Substitutes or Complements?” Journal of Economic Literature 50 2 368-425

Broome J. (1991) “Weighing Goods” Basil Blackwell Oxford

Cowen T. (1989) “Are all Tastes Constant and Identical?” Journal of Economic Behaviour and Organization 11 127-135

Davidson D. (1963) “Actions, Reasons and Causes” Journal of Philosophy 60 23 685-700

Debreu G.(1960) Topological methods in cardinal utility. In K. Arrow, S. Karlin & P. Suppes, editors, Mathematical methods in the social sciences, Stanford University Press, Stanford, California

Fehr E. & Schmidt K.M. (1999) “A Theory of Fairness, Competition and Cooperation” Quarterly Journal of Economics 114 3 817-868

Fodor J. (1989) Psychosemantics: The problem of meaning in the mind Explorations in cognitive science MIT Press Cambridge MA

Frank R. (1988) “Passions within Reason: The Strategic Role of Emotions” W.W. Norton & Company New York

Frey, B.S. & Oberholzer-Gee F. (1997) “The Cost of Price Incentives: An empirical analysis of crowding out” American Economic Review 87, 4 p. 746-55

Geach P.(1957) “Mental Acts: Their Content and their Objects” London Routledge & Kegan Paul.

Geanakoplos J. & Pearce D. (1989) “Psychological Games and Sequential Rationality” Games and Economic Behavior 1, 60-79

Gorman W.M. (1968) “The Structure of Utility Functions” Review of Economic Studies” 35 p. 367-390

Grant R. (2012) “Strings Attached: Untangling the Ethics of Incentives” Russell Sage Foundation and Princeton University Press, Princeton, New Jersey.

Harsanyi J. (1967-1968) “Games with Incomplete Information played by “Bayesian” Players” Parts I-III Management Science 14, 159-182, 320-334, 486-502

Hirsch A.(1977) “Social Limits to Growth” Routledge & Kegan Paul London

Keeney R.L.& Raiffa H. (1976) “Decisions with Multiple Objectives:Preferences and Value Tradeoffs” John Wiley and Sons New York

Lancaster K. (1966) “A New Approach to Consumer Theory” The Journal of Political Economy 74 p. 132-57

Loomes G. & Sugden R. (1982) “Regret Theory: An alternative Theory of Rational Choice under Uncertainty” Economic Journal 92 p. 805-824

Lipman B.L. (1991) “How to Decide How to Decide How to….: Modeling Limited Rationality” Econometrica 59 4 1105-1125

Locke J. (1998) “An Essay Concerning Human Understanding” (first published 1689) Wordsworth editions Limited Hertfordshire

Mansbridge J. “Starting with nothing: on the impossibility of grounding norms in self-interest” in Berner A. & Putterman L. (eds) (1998) “Economics Values and Organization” Cambridge University Press Cambridge.

O’Brien L. & Soteriou M. (eds) (2009) “Mental Actions” Oxford University Press Oxford

Papineau D. (2012) “Philosophical Devices” Oxford University Press, Oxford

Peter (2004) “Choice, Consent and the Legitimacy of Market Transactions” Economics and Philosophy 20 1 1-18

Proust J. (2009) “Is there a Sense of Agency for Thought?” in O’Brien L. & Soteriou M. (eds) (2009) “Mental Actions” Oxford University Press Oxford

Proust J. (2001) “A Plea for Mental Acts” Synthese 129 105-128

Rubinstein A. (1991) “Comments on the Interpretation of Game Theory” Econometrica 59, 4, 909-924

Rubinstein A (1998) “Modeling Bounded Rationality” MIT Press Cambridge Massachussetts

Sanford D.H. (1989) “If P, then Q; Conditionals and the Foundations of Reasoning” Routledge London

Sandel M. (2012) “What Money Can’t Buy: The Moral Limits of Markets” Penguin books London

Smith M. (1987) “The Humean Theory of Motivation” Mind 96 381 36-61

Sperber D. (1996) “Explaining Culture: A Naturalistic Approach” Blackwell, Oxford

Stigler G. & Becker G. (1977) “De Gustibus Non Est Disputandum” American Economic Review

Weirich P. (2010) “Utility and Framing” Synthese 176 83-103

Weirich P. (2001) “Decision Space: Multidimensional Utility Analysis” Cambridge Studies in Probability, Induction and Decision Theory Cambridge Cambridge University Press.

Weirich P. (1980) “Conditional Utility and its place in decision theory” The Journal of Philosophy vol. 77 no. 11p.702-715

Williams B. (2002) “Truth and Truthfulness” Princeton University Press, Princeton, New Jersey

Appendix:

The following proofs require the following simple lemma:

Lemma 1:

If the indicative utility function $U(A(\mathbf{x})\wedge q_i/(A(\mathbf{x})\rightarrow q_i))$ is equivalent to the subjunctive utility function then the utility of the outcome s_i is given by $U^*(q_i)$ where $U^*(\text{-})$ is the subjunctive utility function.

Proof:

From definition of the utility of indicative and subjunctive conditionals:

$U(A(\mathbf{x})\wedge q_i/(A(\mathbf{x})\rightarrow q_i))= U^*(A(\mathbf{x})\wedge q_i\wedge(A(\mathbf{x})\rightarrow q_i))$

Substituting the subjunctive conditional utility:

$U^*(A(\mathbf{x})\wedge q_i\wedge(A(\mathbf{x})\rightarrow q_i))$

$= U^*(q_i\wedge q_i)$ by modus ponens

$=U^*(q_i)$ by tautology

□

Proposition 1: If the subjunctive and indicative conditional utilities are equivalent then the utility of the outcome s_i which occurs after mental action T_j is played can be expressed as:

$$U(A(\boldsymbol{x}) \wedge q_i) = U\big(A(\boldsymbol{x})\big|\forall k\colon \neg(A(\boldsymbol{x}) \rightarrow q_i)\big) + U(A(\boldsymbol{x})|A(\boldsymbol{x}) \rightarrow q_i) + U(q_i|A(\boldsymbol{x}) \rightarrow q_i)$$

when expressed as indicative conditional utilities or:

$$U(A(\boldsymbol{x}) \wedge q_i) = U^*(\forall k\colon A(\boldsymbol{x}) \wedge \neg q_k) + U^*(q_i)$$

when expressed as subjunctive conditional utilities.

Proof:

For all k: $A(\mathbf{x})\rightarrow q_k$ is preferentially independent of $\neg\,[A(\mathbf{x})\rightarrow q_k]$ by assumption.

¬ [A(**x**)→q_k] is preferentially independent of A(**x**)→q_k because ¬ [(A(**x**)→q_k] ↔ A(**x**)∧¬q_k, A(**x**) is preferentially independent of ¬ A(**x**), ¬q_k is preferentially independent of q_k hence by (Gorman 1968 Theorem 1) the conjunction A(**x**)∧¬q_k is preferentially independent of its complement. Looking at ∀k ¬ [A(**x**)→q_k], this is equivalent to ¬ [A(**x**)→q_1] ∧ ¬ [A(**x**)→q_2] ∧ ¬ [A(**x**)→q_3] ∧….∧ ¬ [A(**x**)→q_n]. By repeated use of Gorman's theorem, it follows that this is preferentially independent of its complement.

Hence A(**x**)→q_k is preferentially independent of ¬ [A(**x**)→q_k] for all k and ∀k ¬ [A(**x**)→q_k] is preferentially independent of its complement. Again, using Gorman's theorem, all subsets of this set of predicates are preferentially independent of their complements. This means they are all mutually preferentially independent of each other.

Focus on outcome i: If A(**x**)∧q_i is conditioned on A(**x**)→q_k for each k or ∀k ¬ [A(**x**)→q_k] then we have a utility tree structure. This means that the conditional predicates (A(**x**)∧q_i)/ (A(**x**)→q_k) for each k and (A(**x**)∧q_i)/ ∀k:¬ [A(**x**)→q_k] must be mutually preferentially independent of each other as well (Gorman 1968). It follows that the utilities of these conditional predicates are additively separable (Debreu 1960).

Suppose that for outcomes s_k with k=1….w, s_k does not follow T_j , while for outcomes s_k with k=w+1….i-1,i+1,…..,n , s_k does follow T_j where n is the total number of outcomes in the game.

Hence:

$$U(A(\boldsymbol{x}) \wedge q_i) = U\big(A(\boldsymbol{x}) \wedge q_i \big| \forall k: \neg(A(\boldsymbol{x}) \rightarrow q_k)\big) + U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_1) + \cdots$$
$$+ U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_w) + U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_{w+1}) \ldots$$
$$+ U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_i) + \cdots U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_n)$$

For k=1…w, $U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_k)$ is equal to zero since it is false that A(**x**) holds because T_j cannot have been played and also s_i cannot have been reached.

For k=w+1….i-1,i+1,…..,n $U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_k)$ is also equal to zero. This is because s_i does not have predicates from other outcomes influencing its utility and A(**x**) conditioned on A(**x**)→q_k, (k≠i) is not relevant to A(**x**)∧q_i.

We have assumed that A(**x**) and q_i are additively separable.

Hence:

$$U(A(\boldsymbol{x}) \wedge q_i) = U\big(A(\boldsymbol{x})\big|\forall k{:}\,\neg(A(\boldsymbol{x}) \rightarrow q_k)\big) + U\big(q_i\big|\forall k{:}\,\neg(A(\boldsymbol{x}) \rightarrow q_k)\big) + U(A(\boldsymbol{x})|A(\boldsymbol{x}) \rightarrow q_i) + U(q_i|A(\boldsymbol{x}) \rightarrow q_i)$$

By assumption the utility of q_i is solely determined by $A(\mathbf{x})\rightarrow q_i$. It follows that $U(q_i/\forall k{:}\neg(A(\mathbf{x})\rightarrow q_k))$ has a utility of zero.

Hence:

$$U(A(\boldsymbol{x}) \wedge q_i) = U\big(A(\boldsymbol{x})\big|\forall k{:}\,\neg(A(\boldsymbol{x}) \rightarrow q_k)\big) + U(A(\boldsymbol{x})|A(\boldsymbol{x}) \rightarrow q_i) + U(q_i|A(\boldsymbol{x}) \rightarrow q_i)$$

The last two terms can be reconstituted giving equation X:

$$U(A(\boldsymbol{x}) \wedge q_i) = U\big(A(\boldsymbol{x})\big|\neg(A(\boldsymbol{x}) \rightarrow q_i)\big) + U(A(\boldsymbol{x}) \wedge q_i|A(\boldsymbol{x}) \rightarrow q_i) \ldots\ldots\ldots\ldots (X)$$

Looking at the first term on the RHS of equation X:

$$\forall k{:}\,\neg(A(\boldsymbol{x}) \rightarrow q_k) \vdash \forall k{:}\,A(\boldsymbol{x}) \wedge \neg q_k$$

Hence:

$$U\big(A(\boldsymbol{x})\big|\forall k{:}\,\neg(A(\boldsymbol{x}) \rightarrow q_k)\big) = U(A(\boldsymbol{x})|\forall k{:}\,A(\boldsymbol{x}) \wedge \neg q_k)$$

Assuming that all indicative conditional utilities of A(**x**) are equivalent to subjunctive utilities of A(**x**) then:

$$U(A(\boldsymbol{x})|\forall k{:}\,A(\boldsymbol{x}) \wedge \neg q_k) = U^*(A(\boldsymbol{x}) \wedge \forall k{:}\,A(\boldsymbol{x}) \wedge \neg q_k) = U^*(\forall k{:}\,A(\boldsymbol{x}) \wedge \neg q_k)$$

Looking at the second term on the RHS of equation X, assuming subjunctive and indicative conditional utilities are equal and using Lemma 1:

$$U(A(\boldsymbol{x}) \wedge q_i|A(\boldsymbol{x}) \rightarrow q_i) = U^*(q_i)$$

Hence:

$$U(A(\boldsymbol{x}) \wedge q_i) = U^*(\forall k{:}\,A(\boldsymbol{x}) \wedge \neg q_k) + U^*(q_i)$$

□

Proposition 2: Assuming mutual utility independence between the attribute vectors **x** and **y** and that the prospects:

$(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0})\rightarrow q_i))$ with probability π; $(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i))$ with probability $(1-\pi)$ and

$(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_1})\rightarrow q_i))$ with probability π; $(A(\mathbf{x})\wedge q_i \wedge Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i))$ with probability $(1-\pi)$

are indifferent for all probabilities π then the player will hold the correct representation $A(\mathbf{x})$ and the indicative conditional utility $U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i)$ will be equal to the subjunctive utility $U^*(q_i)$.

Proof:

If **x** and **y** are mutual utility independent of each other then (from Keeney & Raiffa):

$$U^*\left(A(\boldsymbol{x}) \wedge q_i \wedge \left(Q\left(T_j, \boldsymbol{x}, \boldsymbol{y}\right) \rightarrow q_i\right)\right)$$
$$= k_y U(B(\boldsymbol{y}) \wedge q_i | B(\boldsymbol{y}) \rightarrow q_i) + k_x U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_i)$$
$$+ k_{xy} U(B(\boldsymbol{y}) \wedge q_i | B(\boldsymbol{y}) \rightarrow q_i) U(A(\boldsymbol{x}) \wedge q_i | A(\boldsymbol{x}) \rightarrow q_i)$$

Where $k_y = U(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i))$, $k_x = U(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0})\rightarrow q_i))$ and $k_{xy} = 1-k_x-k_y$.

If prospects are indifferent to each other then:

$\pi U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0})\rightarrow q_i))+(1-\pi)\ U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i))$

$= \pi\ U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_1})\rightarrow q_i)) + (1-\pi)\ U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i))$

Collecting terms:

$U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i)) + [U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0})\rightarrow q_i)) - U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i))]\times\pi = U^*(A(\mathbf{x_0}) \wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i)) + [U^*(A(\mathbf{x_1}) \wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_1})\rightarrow q_i)) - U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i))]\times\pi$

Equating coefficients:

$U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i)) = U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i))$

And:

$U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0})\rightarrow q_i)) - U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i)) = U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_1})\rightarrow q_i)) - U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i))$

Normalising:

$U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_0})\rightarrow q_i)) = 0$ and $U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_1})\rightarrow q_i)) = 1$

Hence:

$U^*(A(\mathbf{x_0})\wedge q_i \wedge (Q(T_j,\mathbf{x_0},\mathbf{y_1})\rightarrow q_i)) = k_y = 0$ and $U^*(A(\mathbf{x_1})\wedge q_i \wedge (Q(T_j,\mathbf{x_1},\mathbf{y_0})\rightarrow q_i)) = k_x = 1$

Therefore $k_{xy} = 0$

It follows that:

$U^*(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x},\mathbf{y})\rightarrow q_i)) = U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i)$

The derivation above has reduced the value of $U^*(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x},\mathbf{y})\rightarrow q_i))$ to those parts of the function that are influenced by the vector **x** irrespective of **y**. Hence the only relevant parts of $Q(T_j,\mathbf{x},\mathbf{y})$ are those that are influenced by **x**. Since there are no cross- vector complex predicates this reduces the relevant parts of $Q(T_j,\mathbf{x},\mathbf{y})$ to the subset described by $A(\mathbf{x})$.

It follows that $Q(T_j,\mathbf{x},\mathbf{y}) = A(\mathbf{x})$.

Hence: $U^*(A(\mathbf{x})\wedge q_i \wedge (A(\mathbf{x})\rightarrow q_i)) = U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i)$

By lemma 1 this means that $U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i) = U^*(q_i)$

□

Proposition 3: If two representations defined by the attribute vectors **x** and **y** are mutually utility independent then **x** and **y** are complements if and only if $U(A(\mathbf{x})\wedge q_i / A(\mathbf{x})\rightarrow q_i)$ is monotonically increasing with $U^*(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x},\mathbf{y})\rightarrow q_i))$.

Proof:

For the sake of convenience, utilities will be abbreviated as follows:

$U^*(A(\mathbf{x})\wedge q_i \wedge (Q(T_j,\mathbf{x},\mathbf{y})\rightarrow q_i)) = U^*(\mathbf{x},\mathbf{y})$

$U(A(\mathbf{x}) \wedge q_i / A(\mathbf{x}) \rightarrow q_i) = U(\mathbf{x})$

$U(B(\mathbf{y}) \wedge q_i / B(\mathbf{y}) \rightarrow q_i) = U(\mathbf{y})$

If **x** and **y** are mutually utility independent then we have equation 1:

$$U^*(\boldsymbol{x}, \boldsymbol{y}) = k_x U(\boldsymbol{x}) + k_y U(\boldsymbol{y}) + k_{xy} U(\boldsymbol{x}) U(\boldsymbol{y})$$

Where k_x, $k_y > 0$ and all utilities are positive.

The proof consists of two lemmas:

Lemma A: attribute vectors **x** and **y** are complements if and only if $k_{xy} > 0$.

(This is a formalisation of an informal argument by Keeney & Raiffa p. 240)

Assume two vector examples of **y**, labelled $\mathbf{y_a}$ and $\mathbf{y_b}$ where $\mathbf{y_b} \geq \mathbf{y_a}$ (i.e. for each element $y_{bj} \geq y_{aj}$ but $\mathbf{y_a} \neq \mathbf{y_b}$).

By construction, the vector **y** is monotonic with utility so $U(\mathbf{y_b}) \geq U(\mathbf{y_a})$.

Partially differentiating holding each vector example constant:

$$\left(\frac{\partial U^*(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{\boldsymbol{y}=\boldsymbol{y_a}} = \left(\frac{\partial U^*(\boldsymbol{x}, \boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y}=\boldsymbol{y_a}} \times \left(\frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)_{y=y_a} + \left(\frac{\partial U^*(\boldsymbol{x}, \boldsymbol{y})}{\partial U(\boldsymbol{y})}\right)_{y=y_a} \times \left(\frac{\partial U(\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{y=y_a}$$

And:

$$\left(\frac{\partial U^*(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{\boldsymbol{y}=\boldsymbol{y_b}} = \left(\frac{\partial U^*(\boldsymbol{x}, \boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y}=\boldsymbol{y_b}} \times \left(\frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)_{y=y_b} + \left(\frac{\partial U^*(\boldsymbol{x}, \boldsymbol{y})}{\partial U(\boldsymbol{y})}\right)_{y=y_b} \times \left(\frac{\partial U(\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{y=y_b}$$

Since U(**y**) is independent of **x**:

$$\left(\frac{\partial U(\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{y=y_a} = \left(\frac{\partial U(\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{y=y_b} = 0$$

We can also observe that the variation of **x** in U(**x**) is independent of the level of **y**. Hence:

$$\left(\frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)_{y=y_a} = \left(\frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)_{y=y_b}$$

From this follows equation 2:

$$\left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{\boldsymbol{y=y_a}} - \left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{\boldsymbol{y=y_b}} = \frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}\left[\left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y=y_a}} - \left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y=y_b}}\right]$$

Define the term in square brackets on the RHS as M:

$$M = \left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y=y_a}} - \left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y=y_b}}$$

Keeney and Raiffa (1976) point out that if :

$$\left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{\boldsymbol{y=y_a}} < \left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}\right)_{\boldsymbol{y=y_b}}$$

Then **x** and **y** are complements.

By construction $\frac{\partial U(\boldsymbol{x})}{\partial \boldsymbol{x}}$ is positive so the complementarity of **x** and **y** depends on the sign of M being negative. Differentiating each term in M using equation 1:

$$\left(\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial U(\boldsymbol{x})}\right)_{\boldsymbol{y=y_a}} = k_x + k_{xy} \times U(\boldsymbol{y}_a)$$

Differentiating similarly for $\mathbf{y_b}$ and substituting into M we get:

$$M = k_{xy}\big(U(\boldsymbol{y_a}) - U(\boldsymbol{y}_b)\big)$$

Since U($\mathbf{y_a}$) – U($\mathbf{y_b}$) is negative it follows that for M to be negative $k_{xy} > 0$.

Conversely, if $k_{xy}>0$ then M must be negative and so it follows that the LHS of equation 2 must be negative and so **x** and **y** are complements.

Lemma B:

U(**x**) is monotonic with U***(**x**,**y**) if and only if $k_{xy}>0$.

Given Equation (1), one can differentiate to get equation 3:

$$\frac{\partial U^*(\boldsymbol{x},\boldsymbol{y})}{\partial U(\boldsymbol{x})} = k_x + k_{xy}U(\boldsymbol{y})$$

For U(**x**) to be monotonic with U*(**x**,**y**) then the RHS of equation 3 must be positive. k_x is positive by assumption as is the utility U(**y**). For equation 3 to be positive it must be the case that $k_{xy}>0$.

Conversely, if $k_{xy}>0$ then, given the positivity of k_x and U(**y**), equation (3) must be positive as well and U(**x**) must be monotonic with U(**x**,**y**).

Proof of theorem: Lemma A and Lemma B must hold at the same time when $k_{xy}>0$. Hence the proposition follows.

□